Study on Chemical Composition Subdivision of Glass Relics Based on Random Forest

Jinlong Li[1,2,3(✉)], Jinde Li[1,2,3], and Kejing Chen[1,2,3]

[1] Southwest University, Chongqing, China
ljlsteven@163.com
[2] Shenzhen MSU-BIT University, Shenzhen, China
[3] Chongqing Technology and Business University, Chongqing, China

Abstract. Random Forests is a statistical learning theory-based combinatorial classifier, an integrated learning algorithm based on decision trees, with high prediction accuracy, which combines bootstrap resampling method and decision tree algorithm, the essence of the algorithm is to construct a set of tree classifiers h k (x), k1, and then use the set to classify and predict by voting for classification and prediction. Ancient glass was made locally by absorbing its technology, but the chemical composition was different. In this paper, we used the random forest algorithm to draw a rose diagram of the importance of chemical components affecting the classification pattern of glass, combined the top five chemical components in the rose diagram, and used a hierarchical clustering algorithm to classify them according to their average content, and gave the classification results. The study shows that the random forest algorithm has a good discriminative effect on the classification study of chemical components of glass classification laws, and provides a fast and feasible method for glass classification laws.

Keywords: Random forest · decision tree · classification law · feature selection · subclass classification

1 Introduction

The random forest model is a classification algorithm proposed by Breiman and Cutler in 2001, which improves the prediction accuracy of the model by aggregating a large number of classification trees. Compared with other classic machine learning models, random forest facilitates the calculation of nonlinear effects of variables and can reflect the interaction between variables. In 2008, Anita Prinzie and Dirk Van den Poel 25 combined the two methods, MultiNomial Logit and Random Forest, to implement a new model of multiple classifiers, Random MultiNomial Logit (RMNL). The MultiNomial Logit method, which has drawbacks in handling large feature spaces, is compensated by using Random Forest. This application is another meaningful extension of their proposed framework. Numerous studies on Random Forest have also focused on practical applications, such as spatial feature recognition, complete network detection, and information extraction biochips, which have yielded very promising results. In this paper, we

P. Qi and Z. Chen (Eds.): ICBDIE 2023, AISR 178, pp. 647–653, 2024.
https://doi.org/10.2991/978-94-6463-238-5_85

analyze the classification laws of high potassium glass and lead-barium glass through relevant data; for each glass type, we select the appropriate chemical composition and classify them into subclasses, provide specific classification methods and results, and attempt to analyze the rationality and sensitivity of the classification results.

2 Model Analysis

Firstly, the given data were divided into two categories according to glass types, and the graphs of the frequency of ornamentation and the percentage of color frequency of the two glasses were drawn respectively, and then the graphs were analyzed directly to compare the classification laws of the two glasses due to color and ornamentation; then the random forest indicator algorithm was used and the importance rose diagrams of chemical components affecting the classification laws of glasses were drawn; the top five chemical components in the importance ranking according to the rose diagrams and according to their Then the PAM clustering algorithm and K-means clustering algorithm were introduced, and their clustering rationality was compared according to CH index, RI index and contour coefficient, and sensitivity analysis was again performed on the number of clusters of subclassification of each of the three clustering algorithms regarding the two glasses.

3 Model Building and Solving

3.1 Regular Classification Based on Random Forest Indicator Algorithm

The random forest indicator algorithm was used to analyze the classification patterns of high potassium glass and lead-barium glass based on various chemical compositions of the artifacts. Considering that random forest is generally suitable for large sample data, the training ability needs to be evaluated for data sets with small samples. Therefore, high potassium glass samples and lead-barium glass samples are designated as training data and test data, respectively. It was verified that the random forest can rely on the data of chemical composition can accurately distinguish between high potassium glass and lead-barium glass with 100% correct classification rate, so its performance is excellent on the current sample and can be used as the analysis of small sample data in this paper. In random forest, the m training samples are sampled T times by random sampling with put-back, and each sampling will produce a sample set with sample number m and enter into the parallel T decision trees, and such sampling with put-back will result in some samples in the training set not entering the sample set of the decision trees, and these uncollected samples are the out-of-bag data [4].

Random forest indicator algorithm steps. Step1: Run with out-of-bag data samples on the trained decision tree T_1 You can get the out-of-bag data error ei Step2: Then keep the other columns unchanged and randomly transform the eigenvalues of the column i features or add noise on the eigenvalues of the column i features in the out-of-bag data to get the out-of-bag data error $ei2$ [2]. Step3: Calculate the importance of feature $X = \sum(\text{ei2} - \text{e1})/T$ Step4: Randomly transform or add noise to each feature separately using out-of-bag data, and then iterate to evaluate X after changing each feature [1].

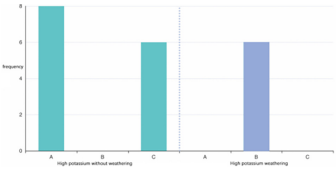

Fig. 1. Frequency high potassium glass ornamentation

Step5: The X is normalized and then ranked, and the higher the value of the index is more important to the classification result. For the classification law of high potassium glass and lead-barium glass, the frequency of weathering based on the relevant variables can be counted based on the original data. Considering the small effect of weathering on the type of glass samples, the sample data are similarly divided into two categories: high potassium glass and lead-barium glass. Figures 1, 2, 3, and 4 shows the frequency of weathering status of ornamentation and color for different types of glass.

The percentage of different types of glass samples was statistically calculated, for the decoration of different types of glass samples, only A and C decoration existed in non-weathered state in high potassium glass, and only B decoration existed in weathered state; no B decoration existed in lead-barium glass, and the percentage of A and C decoration

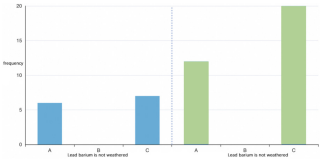

Fig. 2. Frequency lead-barium glass ornamentals

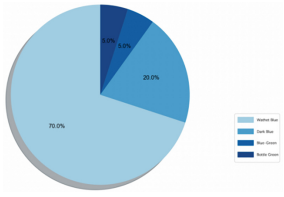

Fig. 3. Color frequency proportion of high potassium

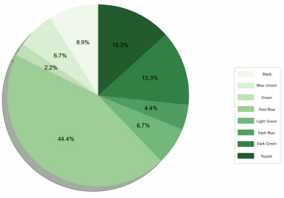

Fig. 4. GlassColor frequency of lead-barium glass

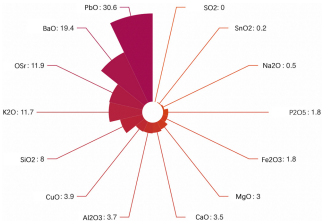

Fig. 5. Importance division of chemical components in glass samples

in non-weathered state was lower than that of high potassium glass; and for the color of different types of glass samples, only four colors existed in high potassium glass, such as blue-green and light blue, and the percentage of blue-green reached 70%, and lead-barium glass is present in all glass colors, of which light blue is predominant, accounting for about 4.4%. The mean importance rose plot of each oxide (hereafter referred to as classification factor) in the glass samples obtained by data normalization using the random forest indicator algorithm based on out-of-bag data error rate is presented. As shown in Fig. 5.

The importance of barium oxide and lead oxide is higher in different types of glass samples, and the importance of the factors is about 20%–30%, followed by strontium oxide, potassium oxide and silica. The first five oxides mainly include the unique oxides with high content in high potassium glass and lead-barium glass, and their compositions are more relevant to the classification pattern.

3.2 Classification of Ancient Glass Subclasses Based on Hierarchical Clustering

Hierarchical Clustering (HC) creates a hierarchical nested clustering tree by calculating the similarity between different types of data points. In a clustering tree, different types of original data points are the bottom level of the tree, and the top level of the tree is the root node of a cluster It has two clustering modes, bottom-up type and top-down type, respectively. In order to select appropriate chemical compositions for high-potassium glass and lead-barium glass, subclasses were divided. Firstly, it is necessary to clarify the specific elements for subclass classification, so we choose to start from the importance of different types of classification factors from the previous question, and select the top five classification factors in terms of importance under normalized conditions, i.e., lead oxide, barium oxide, strontium oxide, potassium oxide and silica, whose selected classification factors are representative in glass sample Later, according to the mean content of each oxide in different types of glass samples, the classification was made according to lead-barium glass and high-potassium glass. And separately defined lead barium glass subclass $Pb - BaI$ and high potassium glass subclass KII, specifically expressed as follows.

$$Pb - Ba_I = \begin{cases} 0, \text{High lead type lead barium glass} \\ 1, \text{Lowu strontium type lead barium glass} \\ 2, \text{High barium type lead barium glass} \end{cases}$$

$$K_{II} = \begin{cases} 0, \text{High silicon type high potassium glass} \\ 1, \text{High potassium glass} \end{cases}$$

A top-down type of clustering model was selected, and the different types of glass samples were run separately using Ward's method for iterative calculations, i.e., 2 clusters were continuously selected for merging to make the smallest increase in variance among all clusters.

3.3 Analysis of Classification Results of Subclass Glass Specimens

For the analysis of subclass division results, PAM (Partitioning Around Medoid) algorithm and K-means clustering (K-means) algorithm are introduced to compare and analyze with hierarchical clustering. The steps of PAM algorithm are as follows: the whole sample is X and k clustering centers are randomly selected in the sample. The distance to each cluster center is calculated for the sample points other than the cluster center, and then the sample is grouped into the sample points closest to the sample center, and the minimum value of the sum of the distances to all other points is calculated for the sample points in each class except for the point in the center. The minimum value is used as the new cluster center to achieve a clustering optimization. The previous step is repeated until the positions of the cluster centers no longer change twice. Rationalization analysis of the subcategory classification results: The external metric CH and internal metric RI with contour coefficients were selected to analyze the reasonableness of the classification results, and their metrics were introduced as follows. The CH index measures the internal tightness of classes by calculating the sum of squares of the distances between each point in the class and the center of the class, and the separation of the data set by calculating the sum of squares of the distances between the center of each class and the center of the data set; the CH index is the ratio of separation to tightness, so the larger the CH value is, the tighter the class itself is and the more dispersed the class is, i.e., the more reasonable the classification result. The RI index is mainly used to evaluate the classification results by applying the principle of permutation and combination. The calculation of the index.

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

The contour coefficient is widely used to evaluate the merits of clustering results: the contour coefficient takes the value of $[-1,1]$, and when it tends to be closer to 1, it means that the cohesion and separation are better, and the classification result is more reasonable. The contour coefficient of the i-th object in the clustering is

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the intra-cluster dissimilarity, which is the average of dissimilarity from i vector to other points in the same cluster, reflecting the cohesion; $b(i)$ is the inter-cluster dissimilarity, which is the minimum of the average dissimilarity from i vector to other clusters, reflecting the separation [3]. Finally, the contour coefficients of all

Table 1. Index statistics of each cluster model

Clustering Model	CH	RI	Contour Coefficient
Hierarchical Clustering	26.110	0.925	0.558
PAM	25.770	0.640	0.440
K-means	25.380	0.115	0.440

points are averaged, which is the total contour coefficient of this clustering result. The summary Table 1 of the indicators related to each subclass of the glass sample based on the multi-analysis method is as in Table 1.

By analyzing the related indexes, it can be obtained that: the hierarchical clustering method is higher than other algorithms in each parameter index, in which the RI value optimization is the most obvious, and the overall model effect: hierarchical clustering PAMK-means. Sensitivity analysis of the number of clusters i for subclass division. By using the Consensus Cluster Plus package built in R language to change the number of clusters k of the classification algorithm in the lead-barium glass group and the high potassium glass group respectively, and try to draw Tracking Plot (Tracking Plot) separately, and then judge the stability of the clustering algorithm according to the color change of different k values, so as to accurately analyze the influence of the change of values on the classification results The degree of impact of the change in value on the classification result. Tracking Plot: The black stripes at the bottom of the image represent each sample, showing the subclasses to which the samples belong when different values of k are taken, and the different color blocks represent different subclasses. If a sample changes its color classification frequently before and after taking different k values, it means that the sample is classified unstable. If there are too many samples with unstable classification, it means that the classification is unstable under that k value.

4 Conclusion

In this paper, we compared the clustering rationality of three clustering models using various evaluation indexes and selected the optimal hierarchical clustering model, which has a high generalizability. Given that the accuracy of the classifier is accurate enough, the later improvement of the model can focus on the optimization of the clustering model; the model relies on the data mainly on the chemical composition of substances, so it is less restricted and highly portable, and can be used for the subclass classification of other items and data classification after modification.

References

1. YUEQI Tang, YINGRAN Liang, JIE-MING Mak, JIN-YAO Lin. Combining random forest and multi-source spatial big data for building function identification[J]. Intelligent Building and Smart City,2022(06):41–45.
2. Xiao Xiangmei, Yu Jian, Lin Zhixing. A network anomaly traffic detection method using random forest[J]. Journal of Sanming College, 2022,39(03):84–91.

3. Peng YF, Chen JH, He C. Comparison and application of machine learning and deep learning based algorithms for evidence-based data extraction in South China Sea[J]. Modern Intelligence,2022,42(02):55–69.
4. Wang ZG, Yan Z, Zhang BC, Rao ZG, Yu LF, Liu J, Gao JF. Identification of colon cancer biomarkers by random forest algorithm based on microarray data (in English)[C]//. Proceedings of the 4th Hubei Anti-Cancer Association Youth Committee Academic Exchange Conference. [publisher unknown],2012:114–128.